Experimental Analysis of Large Belief Networks for Medical Diagnosis

Malcolm Pradhan*, Gregory Provan, Max Henrion
Institute of Decision Systems Research
4984 El Camino Real, Suite 110, Los Altos, CA 94022
*also Section on Medical Informatics, MSOB X-215
Stanford University, CA 94305

We present an experimental analysis of two parameters that are important in knowledge engineering for large belief networks. We conducted the experiments on a network derived from the Internist-1 medical knowledge base. In this network, a generalization of the noisy-OR gate is used to model causal independence for the multivalued variables, and leak probabilities are used to represent the nonspecified causes of intermediate states and findings. We study two network parameters, (1) the parameter governing the assignment of probability values to the network, and (2) the parameter denoting whether the network nodes represent variables with two or more than two values. The experimental results demonstrate that the binary simplification computes diagnoses with similar accuracy to the full multivalued network. We discuss the implications of these parameters, as well other network parameters, for knowledge engineering for medical applications.

ISSUES IN BUILDING LARGE NETWORKS

There is increasing interest in Bayesian belief networks and influence diagrams as representations for medical knowledge that are soundly based on the principles of probability and decision theory. However, questions remain about their practicality for building very large knowledge bases. The work we describe here is part of a long term project to explore and evaluate techniques for knowledge engineering and inference with very large belief networks (BNs). Like any large knowledge-engineering project to encode medical expertise, building a large BN is a lot of effort, and issues of network complexity and the required precision of the probabilities are critical. Presumably, a larger, richer network with more precise probabilities can support more accurate diagnosis. But, what kind of relationships are there between representation and performance?

In this paper, we report some initial experimental results as part of an attempt to help answer these questions. We emphasize that these experiments are not an external validation of the diagnostic accuracy of the network. Here, we focus on two issues: First, what are the effects of alternative ways of expressing the probabilities? We start with frequency integers—0, 1, 2,...,5—to express the links between variables, and compare various mappings from frequencies to conditional probabilities. Second, how much difference does it make to express variables as two-valued or binary (for example, a disease may be present or absent) instead of four-level (for example, a disease may be absent, mild, moderate, or severe). Using four levels should create a more accurate representation and should lead to better diagnosis, but the improvement in performance may not be worth the additional knowledge engineering and computational effort.

Our experimental results demonstrate that the binary simplification computes diagnoses with similar accuracy to the full multivalued network. In addition, the frequency to probability mappings may significantly affect the overall accuracy of the diagnoses.

CPCS: KNOWLEDGE BASE TO BELIEF NETWORK

The Quick Medical Reference–Decision Theoretic (QMR-DT) project seeks to develop practical decision-analytic methods for large knowledge-based systems. The first stage of the project converted the Internist-1 knowledge base [4] (QMR's predecessor) into a binary, two-layered BN [3,12]. In the second stage of the QMR-DT project we are creating a multilayer BN with multivalued variables, and developing efficient inference algorithms for the network.

To create a large multilevel, multivalued BN we took advantage of a rich knowledge base, the Computer-based Patient Case Simulation system, developed over two years by R. Parker and R Miller [7] (CPCS–PM) in the mid-1980s as an experimental extension of the Internist-1 knowledge base. The CPCS–PM system is a knowledge base and simulation program designed to create patient scenarios in the medical sub domain of hepatobiliary disease, for use in medical education. Unlike that of its predecessor Internist-1, the CPCS–PM knowledge base models the pathophysiology of diseases—the intermediate states causally linked between diseases and manifestations. The original CPCS–PM system was developed in FranzLisp. Diseases and intermediate pathophysiological states (IPSs) were represented as Lisp frames [5].

To construct the BN we converted the CPCS–PM knowledge base to CommonLisp and then parsed it to create nodes. We represented diseases and IPSs as four levels of severity in the CPCS BN—absent, mild,

0195-4210/94/$5.00 © 1994 AMIA, Inc.

moderate, and severe. Predisposing factors of a disease or IPS node were represented as that node's predecessors, and findings and symptoms of a disease or IPS node as the successors for that node. In addition to the findings, CPCS contained causal links between disease and IPS frames, we converted these links into arcs in the BN. Frequency weights [11] from the CPCS–PM ranged from 0 to 5 and were mapped to probability values, as described in the next section.

We generated the initial CPCS BN automatically from the knowledge base, we did manual consistency checking using domain knowledge to edit the network. Because the CPCS–PM knowledge base was not designed with probabilistic interpretations in mind, we had to make numerous minor corrections to remove artifactual nodes, to make node values consistent and to confirm that only mutually exclusive values were contained within a node.

As we checked the validity of the resulting network it became clear that in the original CPCS–PM used frequency weights to represent frequencies *and* to control inference in the system. In the initial version of the CPCS BN we have identified, but not corrected, these inconsistencies. We will explore the effects of further knowledge engineering on the performance of the network in future work.

The resultant network has 450 nodes and over 900 arcs. Seventy-four of the nodes in the network are predisposing factors and required prior probabilities; the remaining nodes required *leak* probabilities (described in the *Network Implementation* section) assessed for each of their values. We thus had to assess almost 600 probabilities to specify the network fully.

For our experiments we used a subset of the full network comprising 110 nodes, which is the set of all ancestor and predecessor nodes of three disease nodes—ascending cholangitis, acute viral hepatitis, and alcoholic hepatitis—shown as heavy outlined nodes in Figure 1. Because the complexity of a BN rises exponentially with the size of the network, inference in a sub network can be accomplished in reasonable time, which is not possible for the full CPCS BN.

MEDICAL IMPLICATIONS OF THE NETWORK PARAMETERS

We varied the mapping and domain-size parameters to assess their effect on this BN representation. The first parameter studied was the frequency to probability mapping. In converting the CPCS–PM knowledge base to a BN, we make the assumption that the frequency weights used in the Lisp knowledge base can be mapped to probability values. The default mapping, called *standard*, is based on the interpretation of frequency weights from the original work to convert the Internist-1 knowledge base to a two-level network [11]. The standard mapping used

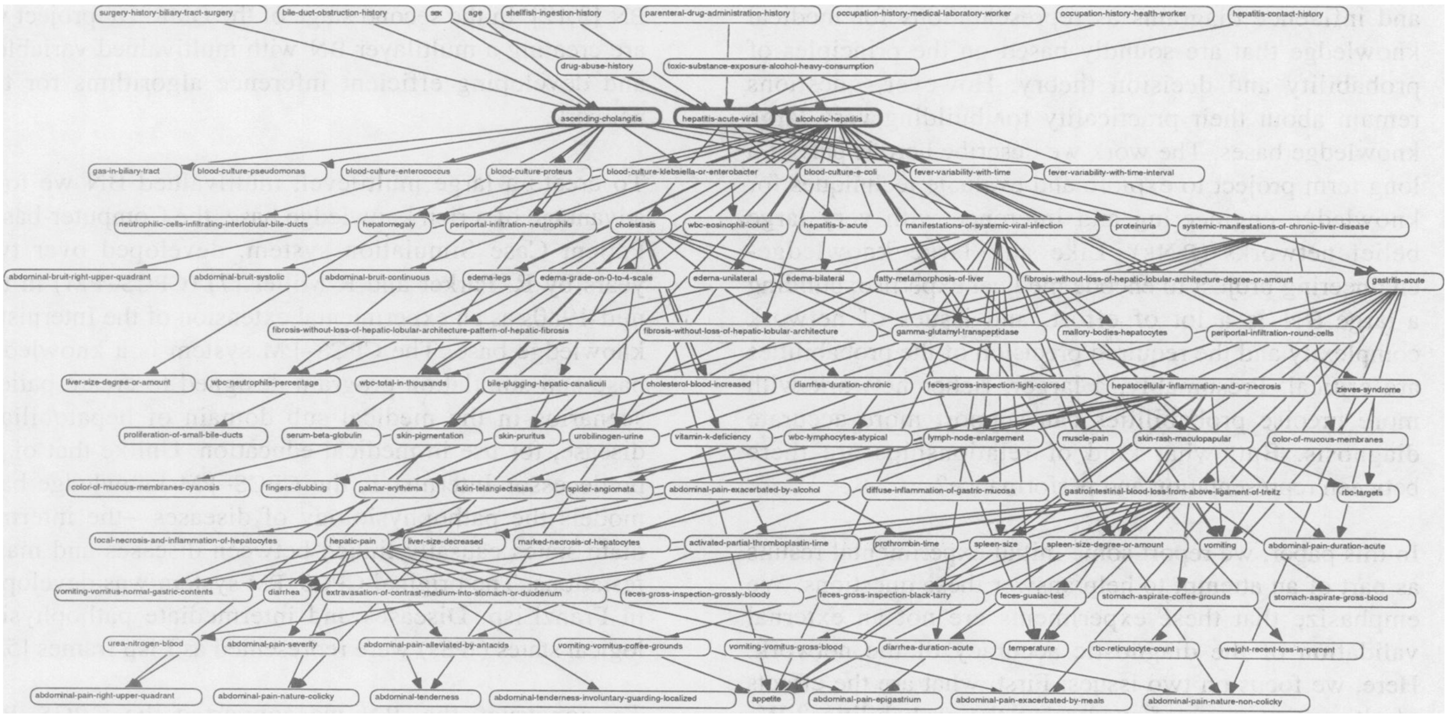
surgery-history-biliary-tract-surgery
bile-duct-obstruction-history
sex
age
shellfish-ingestion-history
parenteral-drug-administration-history
occupation-history-medical-laboratory-worker
occupation-history-health-worker
hepatitis-contact-history
drug-abuse-history
toxic-substance-exposure-alcohol-heavy-consumption
ascending-cholangitis
hepatitis-acute-viral
alcoholic-hepatitis
gas-in-biliary-tract
blood-culture-pseudomonas
blood-culture-enterococcus
blood-culture-proteus
blood-culture-klebsiella-or-enterobacter
blood-culture-e-coli
fever-variability-with-time
fever-variability-with-time-interval
neutrophilic-cells-infiltrating-interlobular-bile-ducts
hepatomegaly
periportal-infiltration-neutrophils
cholestasis
wbc-eosinophil-count
hepatitis-b-acute
manifestations-of-systemic-viral-infection
proteinuria
systemic-manifestations-of-chronic-liver-disease
abdominal-bruit-right-upper-quadrant
abdominal-bruit-systolic
abdominal-bruit-continuous
edema-legs
edema-grade-on-0-to-4-scale
edema-unilateral
edema-bilateral
fatty-metamorphosis-of-liver
fibrosis-without-loss-of-hepatic-lobular-architecture-degree-or-amount
gastritis-acute
fibrosis-without-loss-of-hepatic-lobular-architecture-pattern-of-hepatic-fibrosis
fibrosis-without-loss-of-hepatic-lobular-architecture
gamma-glutamyl-transpeptidase
mallory-bodies-hepatocytes
periportal-infiltration-round-cells
liver-size-degree-or-amount
wbc-neutrophils-percentage
wbc-total-in-thousands
bile-plugging-hepatic-canaliculi
cholesterol-blood-increased
diarrhea-duration-chronic
feces-gross-inspection-light-colored
hepatocellular-inflammation-and-or-necrosis
zieves-syndrome
proliferation-of-small-bile-ducts
serum-beta-globulin
skin-pigmentation
skin-pruritus
urobilinogen-urine
vitamin-k-deficiency
wbc-lymphocytes-atypical
lymph-node-enlargement
muscle-pain
skin-rash-maculopapular
color-of-mucous-membranes
color-of-mucous-membranes-cyanosis
fingers-clubbing
palmar-erythema
skin-telangiectasias
spider-angiomata
abdominal-pain-exacerbated-by-alcohol
diffuse-inflammation-of-gastric-mucosa
gastrointestinal-blood-loss-from-above-ligament-of-treitz
rbc-targets
focal-necrosis-and-inflammation-of-hepatocytes
hepatic-pain
liver-size-decreased
marked-necrosis-of-hepatocytes
activated-partial-thromboplastin-time
prothrombin-time
spleen-size
spleen-size-degree-or-amount
vomiting
abdominal-pain-duration-acute
vomiting-vomitus-normal-gastric-contents
diarrhea
extravasation-of-contrast-medium-into-stomach-or-duodenum
feces-gross-inspection-grossly-bloody
feces-gross-inspection-black-tarry
feces-guaiac-test
stomach-aspirate-coffee-grounds
stomach-aspirate-gross-blood
urea-nitrogen-blood
abdominal-pain-severity
abdominal-pain-alleviated-by-antacids
vomiting-vomitus-coffee-grounds
vomiting-vomitus-gross-blood
diarrhea-duration-acute
temperature
rbc-reticulocyte-count
weight-recent-loss-in-percent
abdominal-pain-right-upper-quadrant
abdominal-pain-nature-colicky
abdominal-tenderness
abdominal-tenderness-involuntary-guarding-localized
appetite
abdominal-pain-epigastrium
abdominal-pain-exacerbated-by-meals
abdominal-pain-nature-non-colicky

Figure 1. We performed experiments on this 110-node subset of the full 450-node CPCS BN. We chose the subset by including all ancestors and predecessors of the disease nodes ascending cholangitis, acute viral hepatitis, and alcoholic hepatitis—shown as dark outlined nodes in row three.

in the two-level BN had a diagnostic performance comparable to the QMR program [3]. We also used two other mappings, *categorical* and *curvilinear*, as shown in Table 1.

There are two reasons to vary the mappings. First, doing so allows us to test whether the interpretation of the standard mappings is accurate. Second, varying the mappings lets us test the sensitivity of a large BN to the probability values. This latter point has an implication for knowledge engineering—when data cannot be found easily it is much easier for experts to assess probabilities in orders of magnitude, say, than as exact values. For example, the categorical mapping interprets the frequency weights as follows: 0 to 3 are small probabilities, and 4 to 5 are high. We empirically determined the cutoff of 3 based on the frequencies already assigned in the network. The curvilinear mapping was determined to be consistent with the frequency values in the network by a domain expert.

Table 1. Mappings used to represent frequency weight from the original CPCS knowledge base as probabilities in the CPCS BN.

	Frequency					
Mapping	0	1	2	3	4	5
standard	0.0025	0.025	0.2	0.5	0.8	0.985
categorical	0.001	0.001	0.001	0.001	0.999	0.999
curvilinear	0.0001	0.001	0.01	0.1	0.9	0.99

The second parameter studied was the domain size of variables. We converted the disease and IPS nodes (which have the four values absent, mild, moderate, and severe) in the original network to nodes with the binary values absent, and present. The domain size influences knowledge engineering: the size of the conditional probability tables grows exponentially with the number of values. The assessment task for experts is also more difficult for multivalued variables compared to binary variables.

NETWORK IMPLEMENTATION

In this section we describe the network used for the experiments, outline the experiments conducted, and discuss the techniques used to analyze the data.

Assumptions

The CPCS network is a multilevel BN in which a *noisy-OR* representation [1,8,9] is used to model the arcs between nodes. The noisy-OR is a simplified BN representation that requires far fewer parameters than does the full conditional probability matrix. The noisy-OR is defined over a set of *binary*-valued variables. Consider an effect variable x, which has n cause variables or predecessors, $d_1,...,d_n$. The noisy-OR can be used when (1) each d_i has a probability p_i of being sufficient to produce the effect in the absence of all other causes, and (2) the probability of each cause being sufficient is independent of the presence of other causes [2].

The noisy-OR gate is commonly used in binary-valued networks to model causal independence. In the CPCS BN, however, disease nodes may have four values (absent, mild, moderate, severe). To accommodate this requirement in the CPCS BN, we use a generalization of the noisy-OR gate called the *noisy-MAX*. Like the noisy-OR, the number of probabilities required to specify the noisy-MAX grows linearly, in contrast to the exponential space requirements of the full specification of conditional probabilities in the network. The specification of a complete conditional probability matrix for a node m with s_m values and n predecessors requires the assessment of $(s_m - 1)\prod_{i=1}^{n} s_i$ probabilities, where s_i is the number of values of predecessor i (for a binary network this reduces to 2^n). In contrast, the causal independence assumption in the form of a noisy-gate reduces this assessment task to $\sum_{i=1}^{n}(s_m - 1)s_i$ probabilities.

Like any other knowledge representation scheme, the BN representation suffers from incompleteness, in that it typically cannot model every possible case. A *leak variable* represents the set of causes that are not modeled explicitly. A *leak probability* is assigned as the probability that the effect will occur in the absence of any of the causes $d_1,...,d_n$ that are modeled explicitly. If the leak variable is modeled explicitly, then it can be treated like any other cause. In this representation the leak node is always assumed to be on; that is, p(*l*=*true*) = 1.0.

Explicitly representing leak nodes in the CPCS BN would almost double the size of the network, so we represent leaks implicitly in the probability tables of the nodes. We developed Netview [10], a graphical tool for visualizing and knowledge engineering belief networks, to facilitate the maintenance and editing of large networks and the associated leak probabilities.

Noisy-MAX implementation

Consider a generalization of the noisy-OR situation in which each variable is allowed to have a finite discrete state space (rather than just a binary state space). This generalization was first proposed by [2], but he did not describe the algorithmic details. In developing this generalization, we assume that we have a set D of predecessor variables $d_1,...,d_n$. Consider first the case where we have a variable x with a subset D_l of D that are present, with the predecessors indexed by $i,j,...,q$.

The variable domains in CPCS BN are all partially ordered, for example, {absent, mild, moderate, severe}, and it turns out that such a partial ordering is necessary for all variable domains. In the remainder of this paper we assume that all variables have ordered domains.

We denote by starred superscripts the state taken by each variable: $i^*, j^*, \ldots, q^*$. The value x^* of variable x is given by $x^* = \max\{i^*, j^*, \ldots, q^*\}$ [2]. In other words, x^* takes on as its value the maximum of the domain values of its predecessors, given that the predecessors are all independent. The unconditional probability with leak node L of maximum value λ and multiple predecessors each with probability η_i is

$$P(x \le x^*) = P(L \le \lambda) \prod_{i:d_i \in D_l} [\eta_i P(x \le x^*) + (1 - \eta_i)]$$

and the unconditional probability is given by

$$P(x = x^*) = P(x \le x^*) - P(x \le x^* - 1).$$

Using this approach, we can compute the value $P(x = x^* \mid D_l)$ in time proportional to the number of predecessors in D_l.

EXPERIMENTAL METHOD

Given the frequency values specified in the original CPCS–PM knowledge base {0,1,2,3,4,5}, we studied the mappings with associated probabilities shown in Table 1.

The second parameter studied was the maximum number of values allowed in variable domains. A binary-valued approximation of 4-ary diseases and IPS nodes was carried out as follows: for four-valued parent-child variable pair (*d*,*x*) and its two-valued counterpart (*d*2,*x*2), we map

1. P(*x*absent| *d*=absent) to P(*x*2=absent| *d*2=absent), and hence P(*x*2=present| *d*2=absent) = 1 -P(*x*2=absent| *d*2=present).
2. P(*x*2=absent| *d*2=present) as 1 - 1/3{P(*x*=absent| *d*=mild) + P(*x*=absent| *d*=mod) + P(*x*=absent| *d*= severe)}.

These variations resulted in six sub networks derived from the original CPCS BN. We ran a suite of test cases on each of the six sub networks.

Test Cases: We generated test cases for the network by simulation in the QMR knowledge base. We generated 10 cases for each disease, resulting in 30 test cases. The terms in the test cases were mapped to the CPCS network. Because both are derived from the Internist-1 knowledge base, there was a good correspondence between the two terminologies. Findings not present in the CPCS BN were not included in the analysis. When we had set as evidence he findings from the test cases, we recorded the posterior probabilities for the 3 disease nodes after we did inference on the networks.

Table 3. The average posterior probability and 95% confidence interval for the true diagnoses and the misdiagnoses.

Network	Average Diagnosis probability	Average Misdiagnosis probability
standard n–ary	0.9458±0.0688	0.2357±0.0784
standard binary	0.9907±0.0120	0.2419±0.0723
categorical n–ary	0.7433±0.1439	0.0364±0.0246
categorical binary	0.8017±0.1417	0.0198±0.0122
curvilinear n–ary	0.7610±0.1419	0.0277±0.0134
curvilinear binary	0.8090±0.1367	0.0209±0.0093

The QMR simulation mode limited the number of test cases we were able to use because it generated cases with overlapping findings. We attempted to acquire cases from published clinico-pathological conferences but the limited domain of our selected sub network was a constraint.

RESULTS

Using multiple comparison [6] we found no statistically significant difference between the categorical or curvilinear mappings, but there was a significant difference between these mappings and the standard mapping, as shown in Table 3. The average probability for the correct diagnosis is higher in the networks with the standard mapping , but the probability assigned to the incorrect diagnoses (false-postives) is also higher. The curvilinear and categorical mappings show a much lower misdiagnosis rate than the standard mapping.

A two-sample t-test [6] of the binary and n–ary networks revealed no statistically significant difference for the standard (95% confidence interval -0.156, 0.117), categorical (95% confidence interval -0.135, 0.118), or curvilinear (95% confidence interval -0.137, 0.115) mappings.

The confidence intervals are approximate because we have assumed normality for results which are bounded by 0 and 1. An alternative method of analysis which may give more accurate intervals is the bootstrap sampling technique [12]. We did not have the resources to carry out bootstrap analysis for this paper.

DISCUSSION

The standard mapping yielded higher posterior probabilities for the correct and incorrect diagnoses compared to the other mappings. The significance of this finding depends on utility assignments when using this network for decision making. Perhaps one explanation for the relatively high misdiagnoses rate for the standard mapping is that it assigns high probability values to intermediate frequencies (2, 3) which are very common in the CPCS–PM, and therefore the CPCS–BN, and may result in greater weights given to findings which should be assigned lower probabilities.

It is, perhaps, surprising that the curvilinear and categorical mappings do so well compared to the standard mapping, given the extreme probability numbers used (for example, none between 0.1 and 0.9). This finding suggests that diagnostic performance in this belief network is not very sensitive to the exact probability numbers.

The statistically insignificant difference in performance between the *n*-ary and binary representations is very interesting. It suggests that the additional effort to develop 4-level instead of binary variables will not be justified by improved diagnosis. The number of probability numbers which need to be specified goes up exponentially with the domain size. For example, for each finding that can be caused by five (n) diseases, you need to specify six (n+1) probabilities for the binary case (Noisy-OR with a leak), compared to 48 (4-1)[1+n(4-1)] probabilities for the four-level case (Noisy-MAX). Hence, the saving in knowledge-engineering effort from using binary instead of 4-level variables is substantial.

These results should be considered as preliminary, for a number of reasons: The test cases are easy, in that they contain a full set of findings, and can be explained by a single disease. In future research, we plan to try harder cases (including phased introduction of findings related to their cost), a more complete network, other mappings, and other sub networks. If the findings hold up in future studies, they could be of substantial practical importance in guiding the development of belief networks with an appropriate balance of effort in knowledge engineering and diagnostic performance. At the very least, these findings should give reassurance to those expert physicians providing expertise to create belief networks who are concerned about the precision with which they can assess subjective probabilities.

ACKNOWLEDGMENTS

This work was supported by NSF Grant Project IRI-9120330, and by computing resources provided by the Stanford University CAMIS project, which is funded by grant number LM05305 from the National Library of Medicine of the National Institutes of Health. The authors thank R.Miller for providing us with access to the CPCS knowledge base, and to Lyn Dupré for her editorial input.

References

[1] Cooper, G. F. A diagnostic method that uses causal knowledge and linear programming in the application of Bayes' formula. *Comp Biomed Res,* 22:223 –237, 1986.

[2] Henrion, M. Practical issues in constructing a Bayes' belief network. In Levitt, T., Lemmer, J. F., and Kanal, L. N. (eds), *Uncertainty in Artificial Intelligence 3*, pages 132–139. North Holland, Amsterdam, 1988.

[3] Middleton, B. et al. Probabilistic diagnosis using a reformulation of the Internist-1/QMR knowledge base-II. Evaluation of diagnostic performance. *Meth Inf Med*, 30:256–67, 1991.

[4] Miller, R. A., Pople, H. E. J., and Myers, J. D. Internist-1: An experimental computer-based diagnostic consultant for general internal medicine. *N Eng J Med*, 307:468–476, 1982.

[5] Minsky, M. A Framework for representing knowledge. In Winston, P. H. (ed), *Psychology of Computer Vision*, pp. MIT Press, Cambridge, MA, 1975.

[6] Ott, R. L. *An Introduction to Statistical Methods and Data Analysis*. Wadsworth, Belmont, CA, 1993.

[7] Parker, R. C. and Miller, R. A. Using causal knowledge to create simulated patient cases: the CPCS project as an extension of Internist-1. *Proceedings of the Eleventh Annual Symposium on Computer Applications in Medical Care,* Los Alamitos, CA, pages 473–480. IEEE Computer Society Press, 1987.

[8] Pearl, J. *Probabilistic reasoning in intelligent systems*. Morgan Kaufman, San Mateo, Ca., 1988.

[9] Peng, Y. and Reggia, J. A. A probabilistic causal model for diagnostic problem solving - Part I: Integrating symbolic causal inference with numeric probabilistic inference. *IEEE Trans SMC*, SMC-17(2):146–162., 1987.

[10] Pradhan, M. et al. Knowledge engineering for large belief networks. *Uncertainty in Artificial Intelligence,* Seattle, Washington, pages 484–490. Morgan Kaufmann, 1994.

[11] Shwe, M. A. et al. Probabilistic diagnosis using a reformulation of the Internist-1/QMR knowledge base-I. The probabilistic model and inference algorithms. *Methods of Information in Medicine*, 30:241–55, 1991.

[12] Tibshirani, R. and Efron. B. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat Sci*, 1:54-77.